

## SCP: Spatial Causal Prediction in Video

Yanguang Zhao<sup>1</sup> Jie Yang<sup>1</sup> Shengqiong Wu<sup>1\*</sup> Shutong Hu<sup>1</sup> Hongbo Qiu<sup>2</sup> Yu Wang<sup>3</sup>  
 Guijia Zhang<sup>2</sup> Tan Kai Ze<sup>1</sup> Hao Fei<sup>1</sup> Chia-Wen Lin<sup>4</sup> Mong-Li Lee<sup>1</sup> Wynne Hsu<sup>1</sup>  
<sup>1</sup>National University of Singapore <sup>2</sup>Shenzhen University  
<sup>3</sup>Sichuan University <sup>4</sup>National Tsing Hua University  
 yanguangzhao@u.nus.edu swu@u.nus.edu

### Abstract

*Spatial reasoning, the ability to understand spatial relations, causality, and dynamic evolution, is central to human intelligence and essential for real-world applications such as autonomous driving and robotics. Existing studies, however, primarily assess models on visible spatio-temporal understanding, overlooking their ability to infer unseen past or future spatial states. In this work, we introduce **Spatial Causal Prediction (SCP)**, a new task paradigm that challenges models to reason beyond observation and predict spatial causal outcomes. We further construct **SCP-Bench**, a benchmark comprising 2,500 QA pairs across 1,181 videos spanning diverse viewpoints, scenes, and causal directions, to support systematic evaluation. Through comprehensive experiments on 23 state-of-the-art models, we reveal substantial gaps between human and model performance, limited temporal extrapolation, and weak causal grounding. We further analyze key factors influencing performance and propose perception-enhancement and reasoning-guided strategies toward advancing spatial causal intelligence. The project page is <https://guangstrip.github.io/SCP-Bench/>.*

### 1. Introduction

Spatial reasoning is a fundamental component of human intelligence that underpins the physical world understanding, enabling perception of spatial relations among objects, the grasp of causality and continuity, and the prediction of future spatial changes [8]. Endowing systems with spatial understanding allows them not only to perceive the visible scene but also to understand the underlying structural regularities and dynamical physical laws, an ability crucial for a wide range of real-world applications such as autonomous driving [11, 20] and robotics [10, 41, 68]. Prior research has made significant progress toward this goal, with early work primarily focusing on static spa-

\*Corresponding author: Shengqiong Wu.

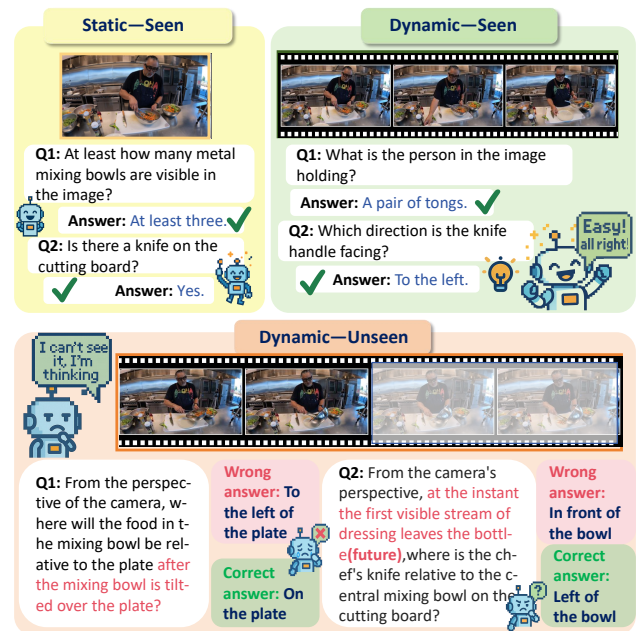


Figure 1. Existing benchmarks primarily assess **known static** or **known dynamic** reasoning based on fully observable scenes. A more challenging **dynamic-unseen** setting is to evaluate models' ability to predict spatial outcomes from partial observations.

tial reasoning that assesses a model's understanding of object layouts and spatial relations from single or multiple viewpoints [15, 27, 46, 55, 57, 59, 65]. More recent work [30, 67, 69] has extended this to spatio-temporal reasoning by introducing videos or dynamic sequences. However, existing efforts [33, 49, 55, 56, 61] remain largely limited to perceptual understanding of observed scenes, assessing whether models can reason about visible spatial attributes (e.g., object size, position, or relations) within given inputs, as shown in Fig. 1. In contrast, real-world spatial reasoning requires the ability to predict future spatial evolution and infer latent causal relations from dynamic observations, beyond merely interpreting what is already visible.

To advance research on causal and temporally aware spatial reasoning beyond the observed scene, we introduce

a new task paradigm, *Spatial Causal Prediction (SCP)*, which addresses the limitations of existing studies confined to visible spatio-temporal understanding. Building on this formulation, we further introduce *SCP-Bench*, a benchmark designed to systematically evaluate a model’s capacity to perceive, reason, and predict spatial causal dynamics under partial observability. We curate various video sources from public datasets and platforms [1, 9, 21, 36], covering diverse **viewpoints** (*egocentric, exocentric, and hybrid perspectives*), and **scenarios** (e.g., *sports, driving, and factory/machine operations*). Such a design allows SCP-Bench to assess not only visual perception but also a model’s grasp of physical commonsense, causal continuity, and dynamical regularities [4, 6, 39, 58]. To construct high-quality annotations, we develop a semi-automatic labeling pipeline combining model-assisted generation [40] with human verification. Following a QA-based task format, we define 8 **spatial reasoning categories** (e.g., *relation, planning, relative size/speed/distance, spatial state*), spanning **two causal directions**, i.e., *backward* (cause inference) and *forward* (result prediction). After rigorous manual validation, the final benchmark comprises **2,500** question–answer (QA) pairs across **1,181** high-quality video clips. With SCP-Bench, we further investigate three key research questions through extensive experiments and in-depth analyses.

#### **RQ1 (§5): How do current MLLMs perform on SCP?**

We evaluate 23 representative state-of-the-art MLLMs [2, 3, 14, 32, 35, 40, 45, 49, 52–54, 62] spanning both open-source and closed-source families under diverse task types, scenes, and settings. Our study yields several striking observations. **First**, models [35, 49] specifically trained for spatial perception often underperform compared to general-purpose models without such specialization, exposing a mismatch between supervised spatial learning and causal generalization. **Second**, the performance gap between long-term and short-term spatial causal prediction is surprisingly small, indicating that current architectures offer limited temporal extrapolation benefits. **Third**, compared with human reasoning, existing models fall short by roughly 22.37% in unseen spatio-temporal causal prediction, underscoring the significant gap in causal understanding. **Fourth**, large open-source models with increased parameter scales achieve performance comparable to several closed-source systems on SCP-Bench, reflecting the rapid advancement of publicly available spatial reasoning models.

#### **RQ2 (§6): Then, what affects their performance?**

To answer this, we design a series of probing tasks to analyze characteristic error patterns and reasoning behaviors. Our results show that current models struggle not only with the spatial perception of visible content but also with predicting unobserved spatial states. On the *perception* side,

understanding dynamic video content remains notably more difficult than handling static visual inputs. Variations in *causal structure, motion dynamics*, and the *continuous evolution of spatial states* often disrupt spatial consistency and result in fragile reasoning [19, 63]. On the *prediction* side, many models lack a solid grounding in physical commonsense, which substantially hinders accurate forecasting of unseen spatial configurations [7, 13].

#### **RQ3 (§7): How to improve their SCP capabilities?**

We further investigate several approaches and paradigms for SCP models. Our experiments reveal that significantly scaling model size consistently yields substantial performance gains. Moreover, we demonstrate that enriching spatial perception with intermediate representations, such as dense video captions or spatial-interaction graphs, provides the predictor with more enriched scene information, leading to slightly improved SCP performance. In addition, incorporating prior knowledge of physical commonsense derived from large language models [40] or learned world models [43] significantly boosts prediction for unseen spatial states. Interestingly, both explicit chain-of-thought (CoT) [17, 48] prompting and implicit self-thinking [22, 25, 47] strategies offer limited improvements.

Overall, this work contributes by **1** for the first time formulating the Spatial Causal Prediction (SCP) task that advances spatial intelligence from observed spatio-temporal understanding to inference over unseen past and future. **2** We develop SCP-Bench, a high-quality benchmark designed to rigorously evaluate spatial reasoning across two causal directions, diverse scenes and viewpoints, and 8 spatial question categories. **3** We perform extensive evaluation and in-depth analysis of existing models on SCP-Bench, uncovering when and why they fail and how to improve them, thereby paving the way for follow-up research.

## 2. Related Work

**Spatial-aware MLLMs.** Recent multimodal large language models (MLLMs) [2, 5, 51] integrate the reasoning capability of large language models [5, 18, 42] with the perceptual strength of vision encoders [24, 34, 37, 50], achieving remarkable progress in visual understanding and cross-modal reasoning [60]. These advances have inspired growing interest in developing world models [23] and embodied agents [16, 31], where spatial understanding plays a central role in grounding models to the physical world. However, achieving robust visual spatial intelligence remains a major challenge, motivating recent efforts toward constructing spatial-aware MLLMs [12, 26, 35, 38, 49, 64]. In contrast to prior studies primarily focusing on spatial reasoning within observed scenes, our work evaluates models’ spatial causal intelligence in unseen videos, aligning more closely with the human ability to infer unobserved spatial dynamics, an



Figure 2. Overview of SCP-Bench. **Left:** Representative examples illustrating the eight task categories. **Right:** Data distribution across scene categories and task types. The benchmark comprises 2,500 QA pairs over 1,181 video clips.

essential capability for real-world embodied applications.

**Benchmarking Spatial Intelligence.** A number of benchmarks assess spatial reasoning in multimodal systems. Early efforts center on 2D imagery (e.g., 3DSR-Bench [33], OmniSpatial [27], Spatial457 [46] and InternSpatial-Bench [15]), while VSI-Bench [55], MMSI-Bench [56] and All-Angles-Bench [57] probe multi-view and relational reasoning, and VLM4D [69] extends evaluation to spatio-temporal video settings. However, these benchmarks largely emphasize static perception or frame-level understanding and thus overlook causal dependencies unfolding over time. More recent dynamic evaluations such as STI-Bench [30] and DSI-Bench [67] incorporate motion, yet remain predominantly perceptual. In contrast, our SCP-Bench explicitly targets forward and backward spatial causal reasoning in dynamic videos, offering a principled testbed for both predictive and diagnostic spatial intelligence. A comprehensive comparison with representative benchmarks is provided in Appendix §E.

### 3. SCP-Bench

**Overview.** We formulate SCP as a multi-choice QA task. Formally, given a video clip with only partial temporal context, along with a question and multiple candidate options, the model aims to select the option that best reflects the underlying spatial causal reasoning. We further introduce SCP-Bench (Fig. 2), a benchmark designed to systematically evaluate this capability of MLLMs on the SCP task. Following, the benchmark design is detailed in Sec. §3.1, and the construction process is described in Sec. §3.2.

#### 3.1. Benchmark Design

To ensure systematic and fine-grained evaluation of SCP, we design SCP-Bench around four complementary dimensions (i.e., *question type*, *causal direction*, *perspective setting*, and *scene diversity*) that together define the benchmark’s structure and scope. We detail these design components below.

**Question Type.** We design 8 task categories to capture variations in spatial causal structure, organized by difficulty. (1) Object-invariant identification (e.g., *Relative Size*) focuses on recognizing the correct entity when object size remains constant. (2) Attribute-dynamic reasoning (e.g., *Appearance Order*, *Relative Speed*, *Spatial State*) probes changes in object attributes under causal evolution. (3) Interaction-level inference (e.g., *Counting*, *Planning*, *Relation*) requires higher-order reasoning over object interactions. Detailed definitions are provided in Appendix §B.

**Causal Direction.** Since the model observes only the visible part before or after the cut point, the task naturally divides into two causal directions: *backward* inference, which reconstructs prior states, and *forward* prediction, which anticipates subsequent scene evolution.

**Perspective Setting.** Since spatial reasoning depends on how a scene is perceived, SCP-Bench includes both single-view and multi-view settings. The single-view setting provides an *ego* (first-person) and *exo* (third-person) perspective. The multi-view setting pairs a video with a reference image from another viewpoint. Specifically, this setting includes *ego-exo* indicating that the video is observed from an ego view but answered from an exo view, with *exo-ego* and *exo-exo* defined analogously.

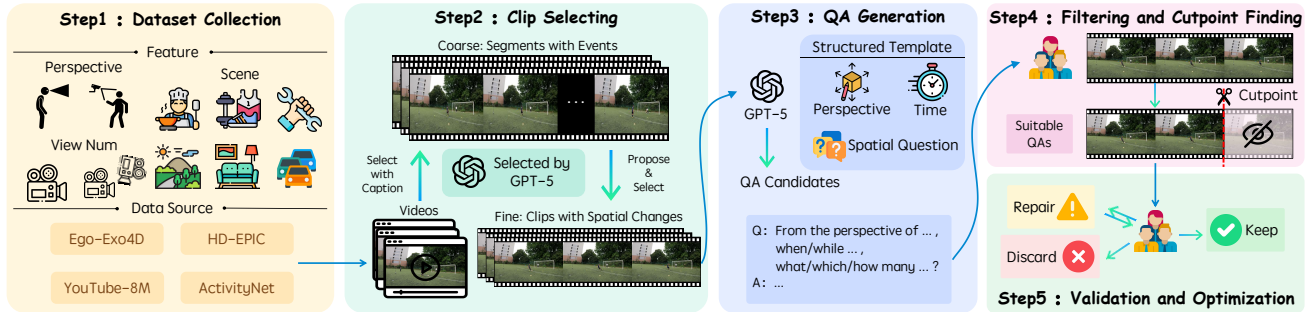


Figure 3. Overview of the SCP-Bench construction pipeline. The process comprises five stages: (1) collection of diverse video sources, (2) clip selection with spatially dynamic segments, (3) generation of candidate QA pairs, (4) QA filtering and cutpoint identification, and (5) dataset validation and refinement.

**Scene Diversity.** To capture the breadth of environments in which spatial causal reasoning occurs, SCP-Bench includes *Artistic Performances*, *Animal Related*, *Sports Related*, *Life Records*, *Factory/Machine Related*, and *Driving-view Related*, ensuring that models are evaluated under diverse visual dynamics and interaction patterns.

### 3.2. Construction Process

We show the detailed construction pipeline in Fig. 3, which comprises the following five key steps:

► **Step-1: Video Source Selection.** To ensure the various attributes of SCP-Bench, we choose the synchronized ego-exo multi-view dataset Ego-Exo4D [21], the high-resolution egocentric dataset HD-EPIC [36], and YouTube-based datasets YouTube-8M [1] and ActivityNet [9].

► **Step-2: Automatic Clip Screening.** Before generating QAs, we apply an automated step to extract high-quality short clips. For each video, GPT-5 [40] expands the provided segments, produces dense temporally aligned captions, and filters for portions that exhibit clear spatial change. It then proposes candidate clips with explicit start-end timestamps, from which we retain the validated ones for downstream QA generation.

► **Step-3: Candidate QA Generation.** To generate challenging and well-controlled QAs, we design a structured prompt template that explicitly encodes three complementary dimensions. The **perspective** fixes the viewpoint from which the question and answer should be interpreted. The **time** dimension specifies the unseen moment or temporal window, outside the visible part, that the model must infer. The **spatial question type** determines which category of spatial reasoning the QA should target. GPT-5 then uses this template to produce diverse and well-structured QA candidates for each selected clip.

► **Step-4: Filtering and Cutpoint Finding.** As our task is highly sensitive to space and time, all QA candidates are manually reviewed. We retain only questions that are unambiguous, answerable from the clip, and genuinely grounded in spatial reasoning, while discarding duplicates or weakly supported items. For each approved QA, annotators first determine a precise cutpoint that separates the clip into visible

and invisible parts, after which GPT-5 generates the corresponding distractor options.

► **Step-5: Validation and Optimization.** To ensure QA quality, annotators conduct multiple rounds of review to refine and correct candidate items. Unanswerable or invalid samples are removed, while fixable issues, such as mismatched question types, unclear perspectives, ill-posed wording or targets, misaligned temporal windows or cutpoints, and defective options, are revised and rechecked. After these, the remaining QAs constitute SCP-Bench.

## 4. Setups of Experiments and Analyses

In the following sections, we fundamentally explore the three tiers of critical questions as raised in Sec. §1. To comprehensively evaluate existing models, we consider a wide range of MLLMs on SCP-Bench, including proprietary, open-source, and spatially specialized models. Specifically, we assess four proprietary models: GPT-5 [40], Gemini-2.5-Pro [14], Gemini-2.5-Flash [14], and Claude-Sonnet-4.5 [3]. For open-source MLLMs, we include Qwen3-VL [54], Qwen3-Omni [53], InternVL3.5 [45], MiniCPM-V-4.5 [62], DeepSeek-VL2 [52], NVILA [32], LLaVA-OneVision [2, 29], and LLaVA-NeXT-Video [66]. Additionally, we evaluate two models specifically designed for spatial reasoning: Spatial-MLLM [49] and SpaceR [35]. All models are tested using their officially recommended configurations to ensure fair and representative results. Each model answers multiple-choice questions based on the input video, question, and answer options, with the accuracy rate serving as the primary evaluation metric. Further implementation details are provided in Appendix §D.

## 5. How Well Do Current Models Perform?

**Overall Evaluation Results.** Table 1 summarizes the overall performance of MLLMs on SCP-Bench. Current systems remain far below human level, underscoring the substantial gap in spatial causal prediction. GPT-5 attains the highest accuracy (66.24%), followed by Qwen3-VL-235B (61.04%) and InternVL3.5-241B (56.96%). Notably, several open-source models match or surpass proprietary ones on specific tasks, for example, outperforming GPT-5 in

Table 1. Evaluation on the SCP-Bench. “Avg.” indicates the overall average accuracy. For each category, the best-performing closed model and open-source model in average score are both indicated in **deep blue**, and best performance on each task is **boxed**.

Model	Avg.	Appearance Order	Counting	Planning	Relation	Relative Distance	Relative Size	Relative Speed	Spatial State
Human Performance	89.61	97.60	81.20	92.26	85.70	86.70	97.62	91.61	84.17
• Closed Models									
GPT-5	<b>66.24</b>	<b>79.04</b>	<b>58.12</b>	<b>59.06</b>	<b>64.07</b>	<b>70.48</b>	<b>95.24</b>	<b>77.42</b>	<b>65.11</b>
Gemini-2.5-Pro	55.84	69.28	54.87	52.76	46.20	63.47	88.10	67.10	62.41
Gemini-2.5-Flash	52.10	59.28	52.14	51.74	43.14	57.75	88.10	66.45	55.60
Claude-Sonnet-4.5	56.14	68.86	52.14	57.43	45.65	60.90	80.95	68.39	63.90
• Open-source Models									
Qwen3-VL-2B	43.04	41.92	42.74	45.01	40.85	44.41	59.52	47.10	40.65
Qwen3-VL-8B	47.52	54.49	51.28	49.29	42.33	49.47	90.48	46.45	46.40
Qwen3-VL-30B-A3B	54.16	65.27	52.14	54.79	46.22	56.65	85.71	66.45	57.19
Qwen3-VL-32B	56.84	59.88	51.28	58.66	52.63	57.98	90.48	67.10	55.04
Qwen3-VL-235B-A22B	<b>61.04</b>	<b>67.07</b>	54.70	60.90	<b>55.03</b>	<b>63.03</b>	<b>97.62</b>	<b>74.84</b>	<b>63.31</b>
Qwen3-Omni-30B-A3B	53.60	63.47	55.56	53.56	47.03	53.72	88.10	65.81	55.40
InternVL3.5-8B	50.52	59.88	54.70	54.79	43.82	54.52	61.90	58.71	44.96
InternVL3.5-38B	53.56	62.28	53.85	56.01	46.34	57.98	90.48	65.81	48.20
InternVL3.5-241B-A28B	<b>56.96</b>	<b>67.07</b>	<b>60.68</b>	<b>61.10</b>	46.11	60.37	90.48	68.39	60.07
MiniCPM-V-4.5	43.80	53.29	49.57	43.99	36.04	49.20	76.19	52.26	42.81
DeepSeek-VL2	38.08	45.51	38.46	39.51	29.41	45.74	73.81	53.55	33.81
NVILA-8B	34.40	36.53	36.75	38.09	30.66	30.05	59.52	38.71	37.05
NVILA-15B	45.28	54.49	45.30	48.07	35.35	52.13	73.81	50.97	49.28
LLaVA-One-Vision-7B	36.48	42.51	37.61	37.07	31.24	38.30	64.29	46.45	35.61
LLaVA-One-Vision-70B	50.84	64.67	52.99	48.68	44.39	53.46	78.57	61.94	51.80
LLaVA-One-Vision-1.5-8B	45.52	56.29	47.01	46.44	39.13	50.27	80.95	51.61	41.73
LLaVA-NeXT-Video-7B	36.60	43.11	25.64	35.44	29.52	48.40	54.76	54.84	32.73
• Spatial Models									
Spatial-MLLM	39.76	45.51	28.21	33.81	38.33	49.73	66.67	50.97	32.37
SpaceR	41.36	52.10	34.19	40.53	34.90	45.21	59.52	54.19	44.60

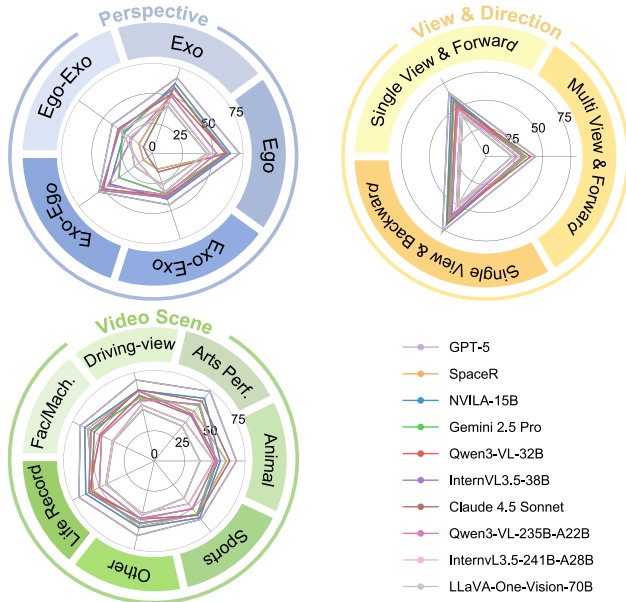


Figure 4. Results across perspectives, view directions, and scenes.

Counting and Planning and achieving comparable results in Relative Size, Relative Speed, and Spatial State. At the level of question types, the difficulty landscape becomes clearer. Relative Size is consistently the easiest, whereas Object Relations, Planning, and Counting are the most challenging, as they require more abstract spatial causal reasoning and higher-order object interaction understanding.

We further examine performance across perspectives,

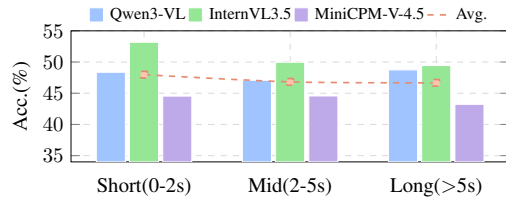


Figure 5. Temporal extrapolation horizon analysis. Samples are grouped by the time gap between the cutpoint and future event: short (0–2s), mid (2–5s), and long (>5s).

causal directions, and scenarios (Fig. 4). Models exhibit clear difficulty with multi-view prediction compared to single-view reasoning, indicating limited perspective correspondence. In causal directionality, models perform better when inferring past (backward) events than future (forward) ones, likely because reasoning from known outcomes is easier than anticipating unseen consequences. Finally, model performance remains relatively balanced across different scene categories, with slightly stronger results in driving-related and factory/machine environments.

**The Impact of Temporal Extrapolation Horizon.** We analyze how model performance varies with different temporal prediction ranges. Specifically, we divide the future duration between the cutpoint and the event completion into three intervals: short (0–2s), mid (2–5s), and long (>5s). We compare three models of comparable size (8B), as shown in Fig. 5. Overall, model accuracy remains relatively stable across horizons, averaging around 46.8%. This indicates that dynamic frame sampling in existing MLLMs

Setting	Qwen3-VL	InternVL3.5	MiniCPM-V-4.5
Base	47.52	50.52	43.80
Gold Video	54.96 $\uparrow$ 7.44	58.64 $\uparrow$ 8.12	43.72 $\downarrow$ 0.08
Caption w/o Video	46.76 $\downarrow$ 0.76	41.24 $\downarrow$ 9.28	45.76 $\uparrow$ 1.96

Table 2. Comparison between perception and reasoning. “Base” is the standard evaluation performance; “Gold Video” evaluates perceptual understanding using unseen part; “Caption w/o Video” tests reasoning based on captions alone. Red and green indicate performance gains and drops, respectively.

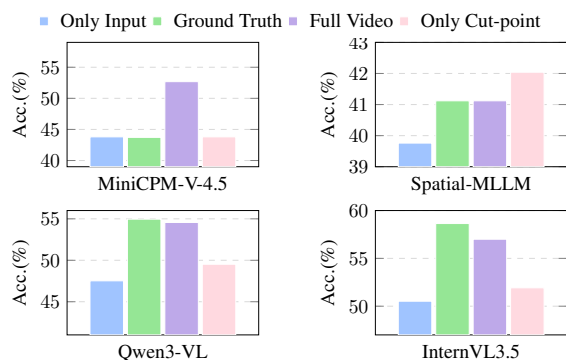


Figure 6. Visible Range Comparison. *Only cutpoint* uses the cutpoint frame as input; *Full video* provides the entire clip; *Ground truth* includes only the unseen clips adjacent to the cutpoint.

mitigates sensitivity to temporal length difference, and also that the current temporal segmentation range may be too narrow to induce significant variation.

**Causal Consistency Evaluation.** To complement the above quantitative results, we conduct a case study to examine whether models obey basic physical and temporal causal constraints. Given a video of a child on a swing, we construct three candidate event progressions after the swing reaches its closest point to the camera: continue forward, stop and reverse, or drift sideways. For each option, the model outputs a rationality judgment (Rationality), a confidence score (Confidence), and an explanation grounded in visible evidence frames under a fixed reference frame. As illustrated in Fig. 8, Qwen3-VL assigns high rationality and confidence to the “continue forward” option, emphasizing local motion continuity while ignoring the physical constraint that the swing should reverse at the turning point. In contrast, GPT-5 explicitly refers to the deceleration, brief stop, and backward motion near the extremum, and correctly rejects this anti-causal trajectory. This case shows that even when models perceive local motion correctly, their explanations can still violate basic causal constraints, revealing a gap in causal consistency for MLLMs.

#### Key Takeaways: Performance in Spatial Causal Prediction

- SCP is a formidable challenge for current MLLMs.
- Large open-source models perform on par with closed ones.
- MLLMs perform similarly across time ranges.

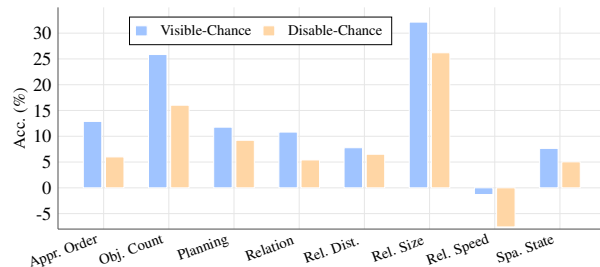


Figure 7. Performance gap comparison. Accuracy improvements over the chance level for vision-enabled (w/ video) and vision-disabled (w/o video) settings across task categories.

Setting	Qwen3-VL	InternVL3.5	MiniCPM-V-4.5	LLaVA-OV-1.5
Base	47.52	50.52	43.80	45.52
w/ Flip	46.76 $\downarrow$ 0.76	48.56 $\downarrow$ 1.96	42.24 $\downarrow$ 1.56	45.32 $\downarrow$ 0.2
w/ CoT	51.52 $\uparrow$ 4.0	50.36 $\downarrow$ 0.16	44.88 $\uparrow$ 1.08	42.88 $\downarrow$ 2.64

Table 3. Comparison with flipped video input and CoT reasoning. “w/ Flip” reverses the input video temporally; “w/ CoT” applies step-by-step reasoning.

## 6. What Affects Spatial Causal Prediction?

**Perception vs. Reasoning Decomposition.** A robust solution to spatial causal prediction requires two essential components: accurate perception of the visual input and reliable reasoning built upon that perception. To investigate which factor plays a more critical role in their poor performance, we design a controlled set of experiments. In one condition, we provide the models with the unseen parts of the clips, referred to as the Gold Video, thereby removing the need for causal inference and turning the task into pure visual understanding. In contrast, we replace the visible parts of the clips with dense captions generated by Tarrier [44], thereby isolating perception and forcing the model to rely solely on textual reasoning. Table 2 shows that the model achieves an average accuracy of 52.44% with the Gold Video input, compared to 44.59% with dense captions, indicating that perception is not the primary bottleneck and that the main limitation lies in spatial causal reasoning.

**Single-Frame vs. Multi-Frame Reasoning.** To examine whether models truly leverage temporal continuity rather than static cues, we compare single-frame and multi-frame reasoning. Specifically, we evaluate model performance when given only the cutpoint frame versus the visible part. As shown in Fig. 6, all four models show a slight improvement over the multi-frame condition when performing single-frame reasoning. This counterintuitive result suggests that temporal cues contribute minimally to model performance under the base spatial causal reasoning task setting, and the observed gains likely stem from static spatial perception rather than genuine temporal understanding.

**Vision Causal Perception.** To further evaluate whether MLLMs can perceive spatial causal logic in the video in-

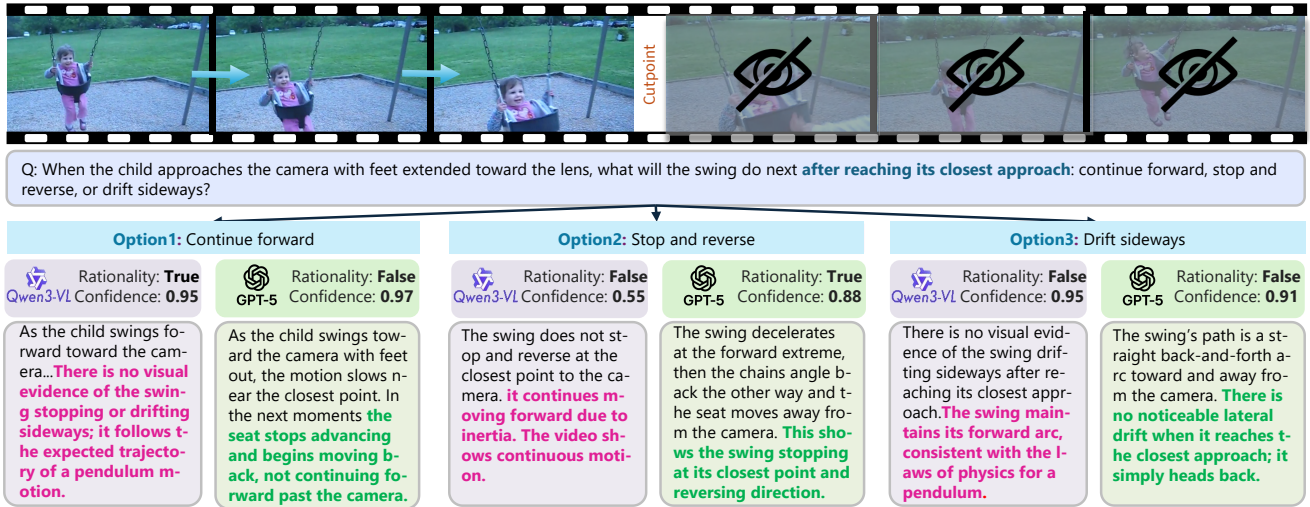


Figure 8. Causal consistency evaluation. Each option shows the model’s reasoning rationality, confidence score, and explanation, assessing whether MLLMs can infer spatial order consistent with physical and causal constraints.

put, we temporally reverse each input clip while keeping the question unchanged and compare performance with the original forward videos. As shown in Table 3, all models show a slight accuracy drop after reversal, with the largest drop being 1.96%. This small drop indicates limited sensitivity to temporal inversion, suggesting that current models have yet to develop a robust notion of spatial causality.

**Visible Range Comparison.** We investigate how the visible range affects model performance. As shown in Fig. 6, models perform best when given the Gold Video input, where direct access to the answer period eliminates the need for spatial causal reasoning and allows performance gains through pure visual understanding. The full-video setting yields slightly lower accuracy but remains above the base condition, suggesting that when the Gold Video part is seen, the visible part introduces mild noise that weakens comprehension. These results indicate that MLLMs perform well when the demand for SCP is minimal, but struggle to infer spatial causality beyond direct observation.

**Is Visual Information Necessary?** As the textual capabilities of models continue to improve, models can sometimes derive correct answers from pure textual reasoning based solely on the knowledge gained during training. This raises the question: can spatial causal reasoning be achieved through pure text alone? We compare model performance with and without video input. The experimental results in Fig. 7 shows that the model’s accuracy drops significantly when no video is provided. Furthermore, as shown in Fig. 11, we attempt to use dense captions generated by Tarsier for the video to replace the video input, the model’s performance still declines. This suggests that although dense captions can partially compensate for the absence of visual input, visual information itself remains indispensable for spatial causal prediction.

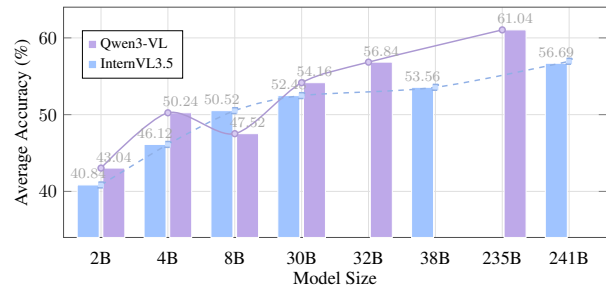


Figure 9. Performance comparison on different model sizes.

🔑 Key Takeaways: What Affects Spatial Causal Prediction

- Perception and Reasoning capabilities limit SCP.
- Visual information is important for SCP.
- MLLMs cannot construct temporal logic to complete reasoning within a video.

## 7. How to Improve Spatial Causal Prediction?

**Model Scale-up Effect Analysis.** We investigate whether enlarging model size leads to improved performance on SCP-Bench by comparing different variants of Qwen3-VL and InternVL3.5 across scales, as depicted in Fig. 9. Overall, performance exhibits a clear upward trend with increased model size, confirming the positive correlation between scaling and spatial reasoning capability. However, the improvement is not strictly monotonic, i.e., small-scale models (e.g., 4B vs. 8B) sometimes show performance fluctuations, suggesting that scaling benefits become more stable and pronounced only when the model grows by a substantial order of magnitude (e.g., 4B → 30B).

**CoT Reasoning.** We further explore the effect of incorporating a vanilla CoT prompt [48] by adding the phrase “think step by step” in the prompt during inference. As shown in Table 3, this simple strategy yields modest im-

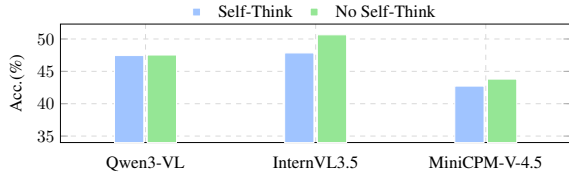


Figure 10. Performance with vs. without self-think reasoning.

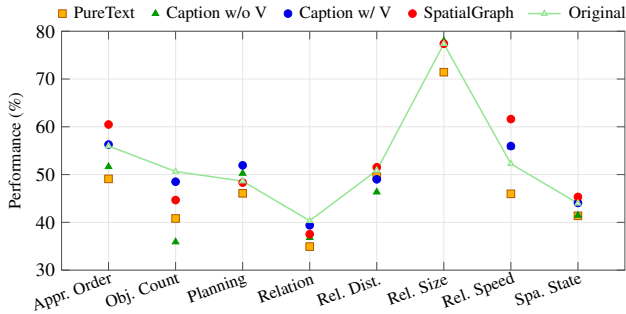


Figure 11. Comparison of perception enhancement strategies. PureText uses only the question; Caption w/ V combines dense video captions with the video and question; Caption w/o V uses only captions with the question; SpatialGraph introduces spatial-interaction graph; Original is the baseline.

improvements for certain models, for example, Qwen3-VL gains over 4% in accuracy. However, the effect is not consistent across architectures; some models show minimal sensitivity or even slight performance degradation, suggesting that the effectiveness of CoT prompting varies with model design and reasoning alignment.

**Self-Think Reasoning.** Recent MLLMs increasingly integrate self-think reasoning through reinforcement learning. To assess the generalizable effectiveness in spatial causal prediction, we compare their performances with and without think-mode reasoning. As shown in Fig. 10, enabling think-mode does not yield consistent improvements; in fact, most models exhibit slight performance degradation, likely due to overextended reasoning chains that introduce noise and divert attention from essential spatial cues.

**Perception Enhancement Strategy.** To address the limited perceptual capability of MLLMs, we explore several mechanisms designed to enhance spatial perception: (1) generating dense captions of the input video clip to enrich scene perception, and (2) constructing spatial-interaction graphs via prompts that capture key objects, environmental elements, and their spatial and interaction relations. The results in Fig. 11 indicate that these perception enhancement strategies lead to only marginal improvements overall, with noticeable gains primarily in specific tasks such as *Appearance Order* and *Relative Speed*. The models fail to leverage spatial-interaction graphs for accurate spatial causal reasoning. When using dense captions to enhance input, the models also exhibit limited benefit, suggesting these perception-level augmentations alone are insufficient to substantially

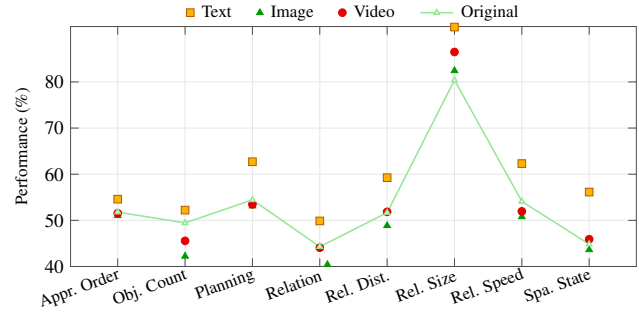


Figure 12. Unseen spatial causal scaffolds for enhanced reasoning. Text provides textual descriptions of future spatial states; Image employs generated future spatial images; Video uses generated future causal videos.

boost spatial causal reasoning.

**Causal Prediction Enhancement Strategy.** We further investigate which types of unseen spatial causal scaffolds are able to effectively enhance reasoning performance. Specifically, we evaluate three forms of auxiliary information: textual descriptions generated by GPT-5 [40], the future spatial images produced by FLUX.1-dev-12B [28], and the future causal videos generated by Wan2.2-TI2V-5B [43]. As shown in Fig. 12, incorporating textual future predictions consistently improves performance across all tasks, likely because MLLMs are inherently more adept at processing and reasoning over textual information. In contrast, image- and video-based scaffolds provide limited gains, likely due to input length constraints, modality noise, and the inherent perception limitations of current MLLMs. Nevertheless, videos outperform images in dynamic-related tasks (e.g., *Relative Size* and *Spatial State*), benefiting from their richer temporal cues.

#### Key Takeaways: How to Improve Spatial Causal Prediction

- *Notably increasing model size helps with SCP.*
- *Unseen spatial causal scaffolds can effectively enhance model performance.*
- *Vanilla CoT and self-thinking mechanisms lead to limited improvements.*

## 8. Conclusion

We introduce Spatial Causal Prediction (SCP) and SCP-Bench, establishing a new paradigm for predictive spatial reasoning beyond visible scenes. Extensive evaluations indicate that current MLLMs remain far from human-level performance, performing better on past inference than future prediction, and relying mainly on perceptual cues. In-depth controlled analyses show that reasoning, rather than perception, constitutes the major bottleneck. While explicit reasoning and structured spatial representations bring limited gains, notably scaling up and integrating causal scaffolds offer a promising path for better SCP performance.

## Acknowledgement

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

## References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 4, 13
- [2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 2, 4, 15
- [3] Anthropic. Claude sonnet 4.5 system card, 2025. Accessed: Apr. 3, 2026. 2, 4, 15
- [4] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 2
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2
- [6] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Proceedings of the NeurIPS*, pages 5083–5094, 2019. 2
- [7] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI*, pages 7432–7439, 2020. 2
- [8] Ruth MJ Byrne and Philip N Johnson-Laird. Spatial reasoning. *Journal of memory and language*, pages 564–575, 1989. 1
- [9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the CVPR*, pages 961–970, 2015. 2, 4, 13
- [10] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *Proceedings of the ICRA*, pages 9490–9498, 2025. 1
- [11] Shitao Chen, Zhiqiang Jian, Yuhao Huang, Yu Chen, Zhuoli Zhou, and Nanning Zheng. Autonomous driving: cognitive construction and situation understanding. *Science China Information Sciences*, 2019. 1
- [12] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Proceedings of NeurIPS*, 2024. 2
- [13] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. 2
- [14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 4, 15
- [15] Nianchen Deng, Lixin Gu, Shenglong Ye, Yanan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025. 1, 3, 21
- [16] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the ACL*, pages 346–355, 2024. 2, 21, 22
- [17] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024. 2
- [18] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 2020. 2
- [19] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 2
- [20] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024. 1
- [21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the CVPR*, pages 19383–19400, 2024. 2, 4, 12
- [22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2
- [23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2
- [24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the CVPR*, pages 16000–16009, 2022. 2
- [25] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 2
- [26] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025. 2

- [27] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025. 1, 3, 21, 22
- [28] Black Forest Labs. Announcing black forest labs, 2024. Accessed: Apr. 3, 2026. 8, 21
- [29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4, 15
- [30] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025. 1, 3, 21, 22
- [31] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 2
- [32] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the CVPR*, pages 4122–4134, 2025. 2, 4, 15
- [33] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the ICCV*, pages 6924–6934, 2025. 1, 3, 21
- [34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [35] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 2, 4, 15
- [36] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the CVPR*, pages 23901–23913, 2025. 2, 4, 12
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pages 8748–8763, 2021. 2
- [38] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 2
- [39] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. 2
- [40] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 2, 4, 8, 15, 21
- [41] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospacial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the CVPR*, pages 15768–15780, 2025. 1
- [42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [43] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 8, 21
- [44] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 6, 15
- [45] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternV3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 4, 15
- [46] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *Proceedings of the CVPR*, pages 24669–24679, 2025. 1, 3, 21, 22
- [47] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 2
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of NeurIPS*, pages 24824–24837, 2022. 2, 7
- [49] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 1, 2, 4, 15
- [50] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024. 2
- [51] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024. 2
- [52] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue

- Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. [2](#), [4](#), [15](#)
- [53] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. [4](#), [15](#)
- [54] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. [2](#), [4](#), [15](#)
- [55] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the CVPR*, pages 10632–10643, 2025. [1](#), [3](#), [21](#), [22](#)
- [56] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. [1](#), [3](#), [21](#), [22](#)
- [57] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. [1](#), [3](#), [21](#), [22](#)
- [58] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. [2](#)
- [59] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshige Yan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV*, 2025. [1](#), [21](#), [22](#)
- [60] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024. [2](#)
- [61] Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zhibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025. [1](#)
- [62] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. [2](#), [4](#), [15](#)
- [63] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the CVPR*, pages 19027–19036, 2023. [2](#)
- [64] Xiaoyu Zhan, Wenxuan Huang, Hao Sun, Xinyu Fu, Changfeng Ma, Shaosheng Cao, Bohan Jia, Shaohui Lin, Zhenfei Yin, Lei Bai, et al. Actial: Activate spatial reasoning ability of multimodal large language models. *arXiv preprint arXiv:2511.01618*, 2025. [2](#)
- [65] Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025. [1](#)
- [66] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. [4](#), [15](#)
- [67] Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. Dsi-bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025. [1](#), [3](#), [21](#), [22](#)
- [68] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. [1](#)
- [69] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the ICCV*, pages 8600–8612, 2025. [1](#), [3](#), [21](#), [22](#)